Rapid alignment methods: FASTA and BLAST

- The biological problem
- Search strategies
- FASTA
- *BLAST*

BLAST: Basic Local Alignment Search Tool

- BLAST (Altschul et al., 1990) and its variants are some of the most common sequence search tools in use
- Roughly, the basic BLAST has three parts:
  - 1. Find *segment pairs* between the query sequence and a database sequence above score threshold (”seed hits”)
  - 2. Extend seed hits into *locally maximal segment pairs*
  - 3. Calculate p-values and a rank ordering of the local alignments
- Gapped BLAST introduced in 1997 allows for gaps in alignments

Finding seed hits

- First, we generate a set of *neighborhood sequences* for given k, *match score matrix and threshold* T
- Neighborhood sequences of a k-word w include all strings of length k that, when aligned against w, have the alignment score at least T
- For instance, let I = GCATCGGC, J = CCATCGCCATCG and k = 5, match score be 1, mismatch score be 0 and T = 4

Finding seed hits

- I = GCATCGGC, J = CCATCGCCATCG, k = 5, match score 1, mismatch score 0, T = 4
- This allows for one mismatch in each k-word
- The neighborhood of the first k-word of I, GCATC, is GCATC and the 15 sequences

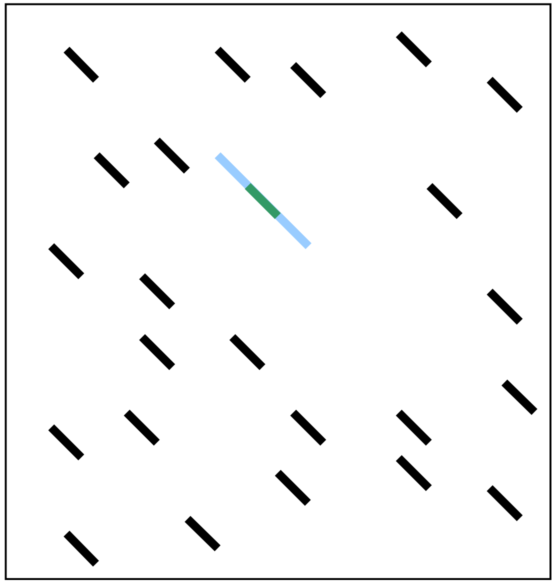
$$\left\{\begin{matrix}A\\C\\T\end{matrix}\right.\mathtt{CATC,G}\left\{\begin{matrix}A\\G\\T\end{matrix}\right.\mathtt{ATC,GC}\left\{\begin{matrix}C\\G\\T\end{matrix}\right.\mathtt{TC,GCA}\left\{\begin{matrix}A\\C\\G\end{matrix}\right.\mathtt{C,GCAT}\left\{\begin{matrix}A\\G\\T\end{matrix}\right.$$

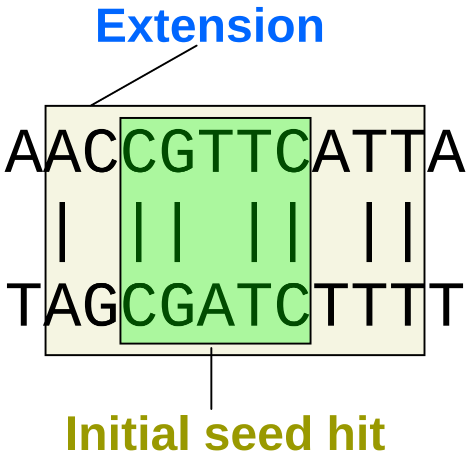
Finding seed hits

- I = GCATCGGC has 4 k-words and thus 4x16 = 64 5-word patterns to locate in J
    - Occurences of patterns in J are called seed hits
- Patterns can be found using exact search in time proportional to the sum of pattern lengths + length of J + number of matches (Aho-Corasick algorithm)
    - Methods for pattern matching are developed on course 58093 String processing algorithms
- Compare this approach to FASTA

Extending seed hits: original BLAST

- Initial seed hits are extended into locally maximal segment pairs or High-scoring Segment Pairs (HSP)
- Extensions do not add gaps to the alignment
- Sequence is extended until the alignment score drops below the maximum attained score minus a threshold parameter value
- All statistically significant HSPs reported

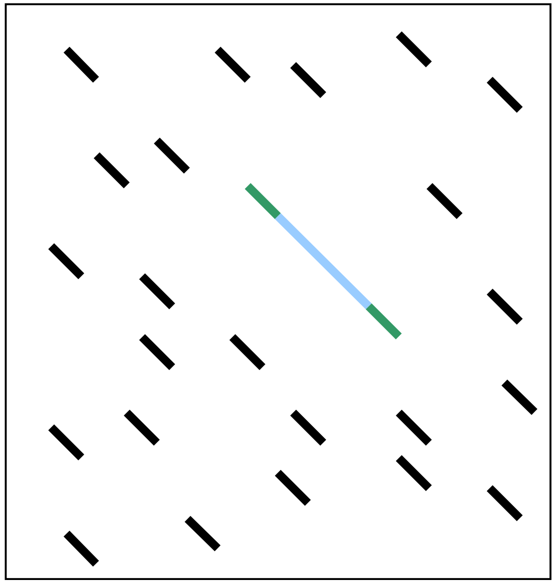




Altschul, S.F., Gish, W., Miller, W., Myers, E. W. and Lipman, D. J., *J. Mol. Biol*., 215, 403-410, 1990

Extending seed hits: gapped BLAST

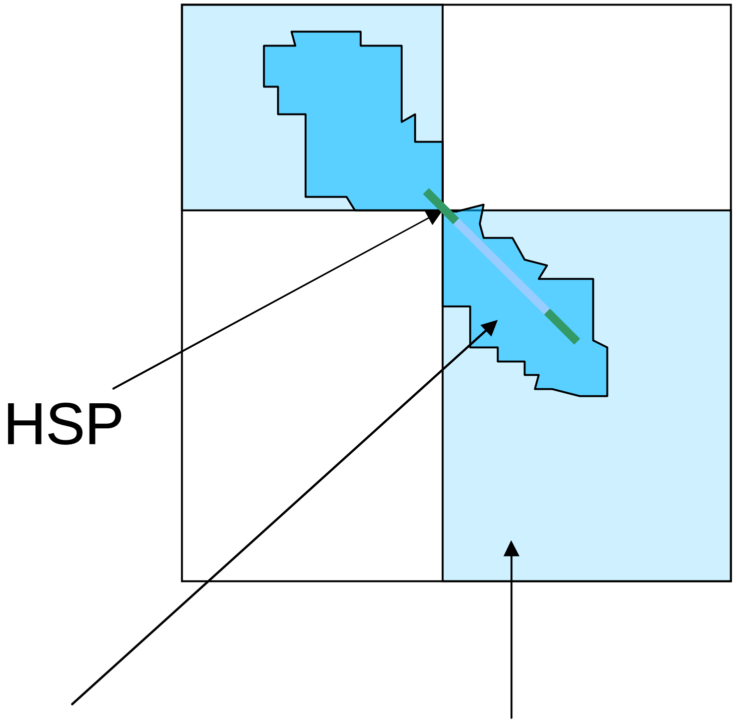
- In a later version of BLAST, two seed hits have to be found on the same diagonal
  - Hits have to be non-overlapping
  - If the hits are closer than A (additional parameter), then they are joined into a HSP
- Threshold value T is lowered to achieve comparable sensitivity
- If the resulting HSP achieves a score at least S_g, a *gapped extension* is triggered



Altschul SF, Madden TL, Schäffer AA, Zhang J, Zhang Z, Miller W, and Lipman DJ, *Nucleic Acids Res*. 1;25(17), 3389-402, 1997

Gapped extensions of HSPs

- Local alignment is performed starting from the HSP
- Dynamic programming matrix filled in ”forward” and ”backward” directions (see figure)
- Skip cells where value would be X_g below the best alignment score found so far



Region searched with score above cutoff parameter

Region potentially searched by the alignment algorithm

Estimating the significance of results

- In general, we have a score $S(D, X) = s$ for a sequence X found in database D
- BLAST rank-orders the sequences found by p-values
- The p-value for this hit is $P(S(D, Y) \geq s)$ where Y is a random sequence
  - Measures the amount of ”surprise” of finding sequence X
- A smaller p-value indicates more significant hit
  - A p-value of 0.1 means that one-tenth of random sequences would have as large score as our result

Estimating the significance of results

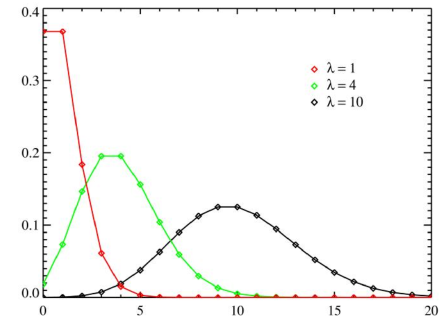
- In BLAST, p-values are computed roughly as follows
- There are nm places to begin an optimal alignment in the n x m alignment matrix
- Optimal alignment is preceded by a mismatch and has t matching (identical) letters
  - (Assume match score 1 and mismatch/indel score -∞)
- Let p = P(two random letters are equal)
- The probability of having a mismatch and then t matches is $(1-p)p^t$

Estimating the significance of results

- We model this event by a Poisson distribution (why?) with mean $\lambda = nm(1-p)p^t$
- P(there is local alignment t or longer)

  $\approx 1 - P(\text{no such event})$

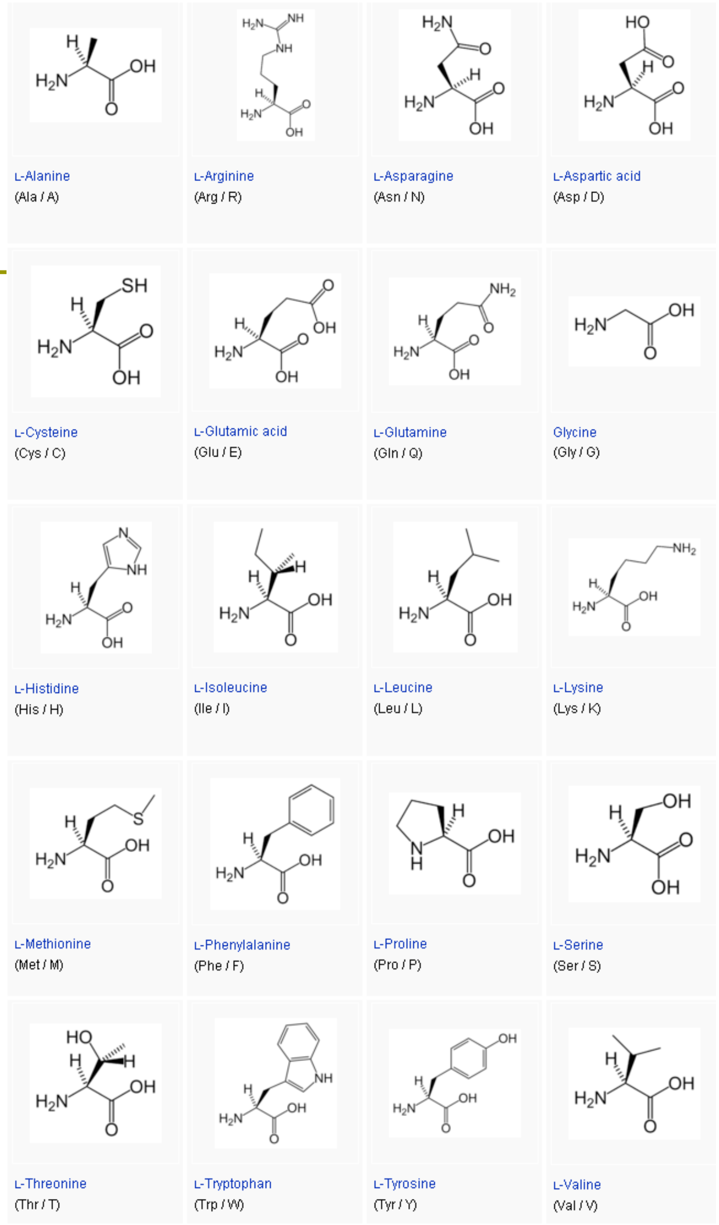
  $= 1 - e^{-\lambda} = 1 - \exp(-nm(1-p)p^t)$



- An equation of the same form is used in Blast:
- E-value = $P(S(D, Y) \geq s) \approx 1 - \exp(-nm\gamma\xi^t)$ where $\gamma > 0$ and $0 < \xi < 1$
- Parameters γ and ξ are estimated from data

Scoring amino acid alignments

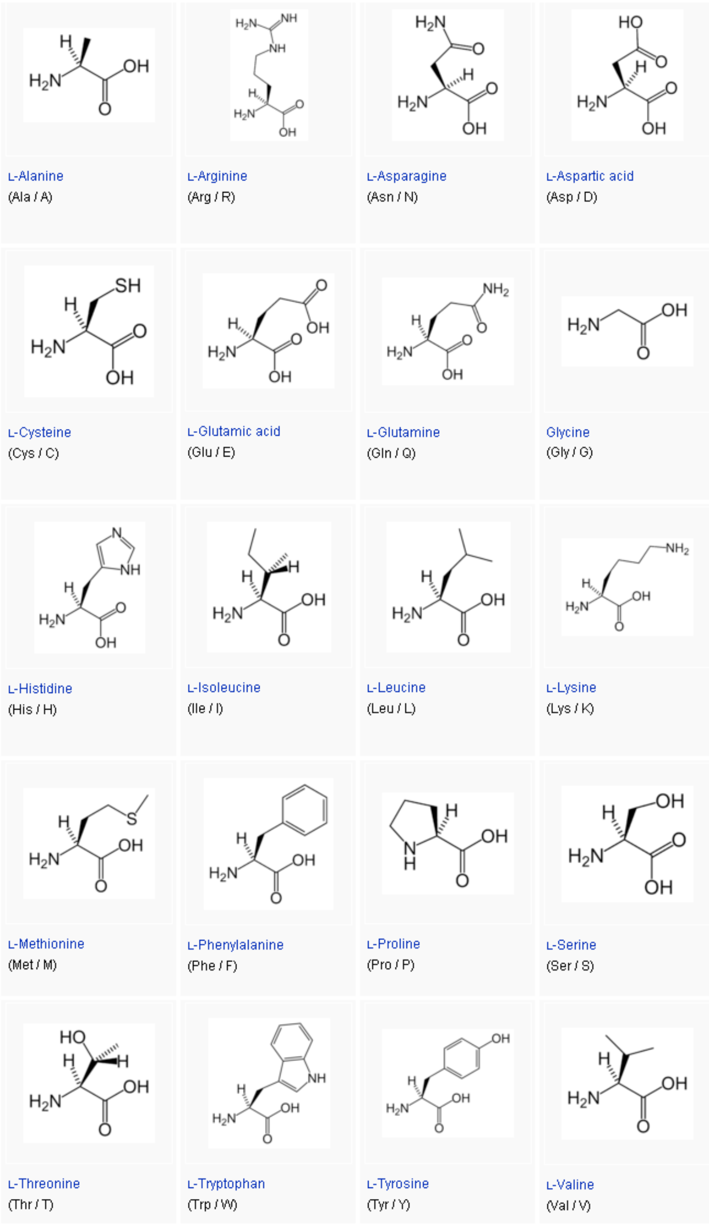
- We need a way to compute the score S(D, X) for aligning the sequence X against database D
- Scoring DNA alignments was discussed previously
- Constructing a scoring model for amino acids is more challenging
  - 20 different amino acids vs. 4 bases
- Figure shows the molecular structures of the 20 amino acids



http://en.wikipedia.org/wiki/List_of_standard_amino_acids

Scoring amino acid alignments

- Substitutions between chemically similar amino acids are more frequent than between dissimilar amino acids
- We can check our scoring model against this



L-Alanine (Ala / A)	L-Arginine (Arg / R)	L-Asparagine (Asn / N)	L-Aspartic acid (Asp / D)
L-Cysteine (Cys / C)	L-Glutamic acid (Glu / E)	L-Glutamine (Gln / Q)	Glycine (Gly / G)
L-Histidine (His / H)	L-Isoleucine (Ile / I)	L-Leucine (Leu / L)	L-Lysine (Lys / K)
L-Methionine (Met / M)	L-Phenylalanine (Phe / F)	L-Proline (Pro / P)	L-Serine (Ser / S)
L-Threonine (Thr / T)	L-Tryptophan (Trp / W)	L-Tyrosine (Tyr / Y)	L-Valine (Val / V)

http://en.wikipedia.org/wiki/List_of_standard_amino_acids

Score matrices

- Scores s = S(D, X) are obtained from score matrices
- Let A = $A_1a_2...a_n$ and B = $b_1b_2...b_n$ be sequences of equal length (no gaps allowed to simplify things)
- To obtain a score for alignment of A and B, where a_i is aligned against b_i, we take the ratio of two probabilities
  - The probability of having A and B where the characters match (match model M)
  - The probability that A and B were chosen randomly (random model R)

Score matrices: random model

- Under the random model, the probability of having X and Y is

$$P(A,B|R) = \prod_i q_{ai} \prod_i q_{bi}$$

where q_{xi} is the probability of occurence of amino acid type x_i

- Position where an amino acid occurs does not affect its type

Score matrices: match model

- Let p_{ab} be the probability of having amino acids of type a and b aligned against each other given they have evolved from the same ancestor c
- The probability is

$$P(A,B|M) = \prod_i p_{a_i b_i}$$

Score matrices: log-odds ratio score

- We obtain the score S by taking the ratio of these two probabilities

$$\frac{P(A,B|M)}{P(A,B|R)} = \frac{\prod_i p_{a_i b_i}}{\prod_i q_{a_i} \prod_i q_{b_i}} = \prod_i \frac{p_{a_i b_i}}{q_{a_i} q_{b_i}}$$

and taking a logarithm of the ratio

$$S = \log_2 \frac{P(A,B|M)}{P(A,B|R)} = \sum_{i=1}^{n} \log_2 \frac{p_{a_i b_i}}{q_{a_i} q_{b_i}} = \sum_{i=1}^{n} s(a_i, b_i)$$

Score matrices: log-odds ratio score

$$S = \log_2 \frac{P(A,B|M)}{P(A,B|R)} = \sum_{i=1}^{n} \log_2 \frac{p_{a_i b_i}}{q_{a_i} q_{b_i}} = \sum_{i=1}^{n} s(a_i, b_i)$$

- The score S is obtained by summing over character pair-specific scores:

$$s(a,b) = \log_2 \frac{p_{ab}}{q_a q_b}$$

- The probabilities q_a and p_{ab} are extracted from data

Calculating score matrices for amino acids

$$s(a,b) = \log_2 \frac{p_{ab}}{q_a q_b}$$

- Probabilities q_a are in principle easy to obtain:
  - Count relative frequencies of every amino acid in a sequence database

Calculating score matrices for amino acids

$$s(a,b) = \log_2 \frac{p_{ab}}{q_a q_b}$$

- To calculate p_{ab} we can use a known pool of aligned sequences
- BLOCKS is a database of highly conserved regions for proteins
- It lists multiply aligned, ungapped and conserved protein segments
- Example from BLOCKS shows genes related to human gene associated with DNA-repair defect xeroderma pigmentosum

Block PR00851A

ID XRODRMPGMNTB; BLOCK
AC PR00851A; distance from previous block=(52,131)
DE Xeroderma pigmentosum group B protein signature
BL adapted; width=21; seqs=8; 99.5%=985; strength=1287

```
XPB_HUMAN|P19447 ( 74)   RPLWVAPDGHIFLEAFSPVYK 54
XPB_MOUSE|P49135 ( 74)   RPLWVAPDGHIFLEAFSPVYK 54
P91579 ( 80)             RPLYLAPDGHIFLESFSPVYK 67
XPB_DROME|Q02870 ( 84)   RPLWVAPNGHVFLESFSPVYK 79
RA25_YEAST|Q00578 ( 131) PLWISPSDGRIILESFSPLAE 100
Q38861 ( 52)             RPLWACADGRIFLETFSPLYK 71
O13768 ( 90)             PLWINPIDGRIILEAFSPLAE 100
O00835 ( 79)             RPIWVCPDGHIFLETFSAIYK 86
```

http://blocks.fhcrc.org

BLOSUM matrix

- BLOSUM is a score matrix for amino acid sequences derived from BLOCKS data
- First, count pairwise matches $f_{x,y}$ for every *amino acid type pair (x, y)*
- For example, for column 3 and amino acids L and W, we find 8 pairwise matches: $f_{L,W} = f_{W,L} = 8$

```
RPLWVAPD
RPLWVAPR
RPLWVAPN
PLWISPSD
RPLWACAD
PLWINPID
RPIWVCPD
```

Creating a BLOSUM matrix

- Probability p_{ab} is obtained by dividing f_{ab} with the total number of pairs (note difference with course book):

$$p_{ab} = f_{ab} / \sum_{x=1}^{20} \sum_{y=1}^{x} f_{xy}$$

- We get probabilities q_a by

$$q_a = \sum_{b=1}^{20} p_{ab}$$

```
RPLWVAPD
RPLWVAPR
RPLWVAPN
PLWISPSD
RPLWACAD
PLWINPID
RPIWVCPD
```

Creating a BLOSUM matrix

- The probabilities p_{ab} and q_a can now be plugged into

$$s(a,b) = \log_2 \frac{p_{ab}}{q_a q_b}$$

  to get a 20 x 20 matrix of scores s(a, b).

- Next slide presents the BLOSUM62 matrix
  - Values scaled by factor of 2 and rounded to integers
  - Additional step required to take into account expected evolutionary distance
  - Described in Deonier’s book in more detail

BLOSUM62

	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V	B	Z	X	*
A	4	-1	-2	-2	0	-1	-1	0	-2	-1	-1	-1	-1	-2	-1	1	0	-3	-2	0	-2	-1	0	-4
R	-1	5	0	-2	-3	1	0	-2	0	-3	-2	2	-1	-3	-2	-1	-1	-3	-2	-3	-1	0	-1	-4
N	-2	0	6	1	-3	0	0	0	1	-3	-3	0	-2	-3	-2	1	0	-4	-2	-3	3	0	-1	-4
D	-2	-2	1	6	-3	0	2	-1	-1	-3	-4	-1	-3	-3	-1	0	-1	-4	-3	-3	4	1	-1	-4
C	0	-3	-3	-3	9	-3	-4	-3	-3	-1	-1	-3	-1	-2	-3	-1	-1	-2	-2	-1	-3	-3	-2	-4
Q	-1	1	0	0	-3	5	2	-2	0	-3	-2	1	0	-3	-1	0	-1	-2	-1	-2	0	3	-1	-4
E	-1	0	0	2	-4	2	5	-2	0	-3	-3	1	-2	-3	-1	0	-1	-3	-2	-2	1	4	-1	-4
G	0	-2	0	-1	-3	-2	-2	6	-2	-4	-4	-2	-3	-3	-2	0	-2	-2	-3	-3	-1	-2	-1	-4
H	-2	0	1	-1	-3	0	0	-2	8	-3	-3	-1	-2	-1	-2	-1	-2	-2	2	-3	0	0	-1	-4
I	-1	-3	-3	-3	-1	-3	-3	-4	-3	4	2	-3	1	0	-3	-2	-1	-3	-1	3	-3	-3	-1	-4
L	-1	-2	-3	-4	-1	-2	-3	-4	-3	2	4	-2	2	0	-3	-2	-1	-2	-1	1	-4	-3	-1	-4
K	-1	2	0	-1	-3	1	1	-2	-1	-3	-2	5	-1	-3	-1	0	-1	-3	-2	-2	0	1	-1	-4
M	-1	-1	-2	-3	-1	0	-2	-3	-2	1	2	-1	5	0	-2	-1	-1	-1	-1	1	-3	-1	-1	-4
F	-2	-3	-3	-3	-2	-3	-3	-3	-1	0	0	-3	0	6	-4	-2	-2	1	3	-1	-3	-3	-1	-4
P	-1	-2	-2	-1	-3	-1	-1	-2	-2	-3	-3	-1	-2	-4	7	-1	-1	-4	-3	-2	-2	-1	-2	-4
S	1	-1	1	0	-1	0	0	0	-1	-2	-2	0	-1	-2	-1	4	1	-3	-2	-2	0	0	0	-4
T	0	-1	0	-1	-1	-1	-1	-2	-2	-1	-1	-1	-1	-2	-1	1	5	-2	-2	0	-1	-1	0	-4
W	-3	-3	-4	-4	-2	-2	-3	-2	-2	-3	-2	-3	-1	1	-4	-3	-2	11	2	-3	-4	-3	-2	-4
Y	-2	-2	-2	-3	-2	-1	-2	-3	2	-1	-1	-2	-1	3	-3	-2	-2	2	7	-1	-3	-2	-1	-4
V	0	-3	-3	-3	-1	-2	-2	-3	-3	3	1	-2	1	-1	-2	-2	0	-3	-1	4	-3	-2	-1	-4
B	-2	-1	3	4	-3	0	1	-1	0	-3	-4	0	-3	-3	-2	0	-1	-4	-3	-3	4	1	-1	-4
Z	-1	0	0	1	-3	3	4	-2	0	-3	-3	1	-1	-3	-1	0	-1	-3	-2	-2	1	4	-1	-4
X	0	-1	-1	-1	-2	-1	-1	-1	-1	-1	-1	-1	-1	-1	-2	0	0	-2	-1	-1	-1	-1	-1	-4
*	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	1

Using BLOSUM62 matrix

```
MQLEANADTSV
 |  | |
LQEQAEAQGEM
```

$$s = \sum_{i=1}^{11} s(a_i, b_i)$$

= 2 + 5 – 3 – 4 + 4 + 0 + 4 + 0 – 2 + 0 + 1

= 7

Demonstration of BLAST at NCBI

- http://www.ncbi.nlm.nih.gov/BLAST/